DOCUMENT RESUME

ED 303 519 TM 012 787

AUTHOR Lombard, Juliana V.
TITLE An Empirical Comparison of a Direct and an Indirect Method of Assessing Writing Proficiency.
INSTITUTION Human Sciences Research Council, Pretoria (South Africa).
REPORT NO ISBN-0-7969-0659-9; RGN-HSRC-P-92
PUB DATE 88
NOTE 62p.; Dutch- and English-language abstracts are provided.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Comparative Analysis; Educational Assessment; *English (Second Language); *Essay Tests; Foreign Countries; Grade 8; Junior High Schools; *Junior High School Students; *Multiple Choice Tests; Objective Tests; Statistical Analysis; *Test Format; Writing (Composition); *Writing Evaluation; Writing Skills
IDENTIFIERS Direct Assessment; *South Africans

ABSTRACT

The validity and reliability of two techniques of assessing writing proficiency were compared in a sample of 300 South African students (in Standard 8) for whom English was a second language. The objective, multiple choice method was compared with a subjective, essay method. Students completed a Standard English Second Language Item Bank Test and wrote an essay and letter. End-of-year examination marks and final report marks were additional criteria. The subjects included 130 boys and 170 girls from schools in Transvaal, Orange Free State, Natal, and Cape Province. Statistical analysis of the four variables--multiple choice test, writing samples, examination grades, and report grades--indicated a substantial correlation between the objective and subjective techniques used, indicating that both approaches tap similar skills. The multiple choice test was found to be a reliable and valid method of testing writing proficiency. Sixteen data tables and a 38-item list of references are provided. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED303519

BEST COPY AVAILABLE

SCOPE OF INTEREST NOTICE

The ERIC Facility has assigned this document for processing to: TM

In our judgment, this document is also of interest to the Clearinghouses noted to the right. Indexing should reflect their special points of view.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. G. GARBERS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

INSTITUTE FOR PSYCHOLOGICAL AND EDUMETRIC RESEARCH P-92

An empirical comparison of a direct and an indirect method of assessing writing proficiency

Juliana V. Lombard

RGN·HSRC

TM021787

An empirical comparison of a direct and an indirect method of assessing writing proficiency

An empirical comparison of a direct and an indirect method of assessing writing proficiency

Juliana V. Lombard

Pretoria
Human Sciences Research Council
1988

Juliana V. Lombard, M.A. (Applied Linguistics), U.E.D.

Institute for Psychological and Edumetric Research
Executive Director: Dr R.J. Prinsloo

ISBN 0 7969 0659 9

© Human Sciences Research Council, 1988

Printed and published by the HSRC
134 Pretorius Street
Pretoria

SYNOPSIS

The aim of this investigation was to examine the validity and reliability of two techniques of assessing writing proficiency. Both measuring techniques, i.e. objective (multiple-choice question) and subjective (essay-type question) have their advantages as well as limitations and little agreement regarding their validity and reliability has been reached to date.

Today great pressure is put on educational bodies to ensure that tests are fair to all those who attempt them and the cry for common standards is steadily increasing. The importance, therefore, of investigating the suitability of a multiple-choice test as a valid and reliable technique of assessing writing proficiency, is self-evident.

The method of investigation involved a study of the relevant literature as well as an empirical study which entailed the application of a Standard 8 English Second Language Item Bank Test and the writing of an essay and a letter by each testee. The testees' end-of-year examination marks as well as their final report marks were obtained from their schools to be used as additional criteria.

A description of the statistical techniques as well as methods to determine the validity and reliability of the four variables (the multiple-choice test, the writing samples, the end-of-year examination marks and the report marks) as used in this study, is rendered.

The results of the statistical analysis indicate a substantial correlation between the objective and subjective measuring techniques involved.

The findings of this investigation strongly suggest that objective and subjective assessments of writing proficiency tend to tap similar skills. Although objective measuring techniques lack face validity and credibility among language teachers and educators generally, the multiple-choice test clearly indicates its suitability as a reliable and valid means of testing writing proficiency.

The possibility of combining objective and subjective assessments into a single measure and reporting only the total score of such a measure may be considered for testing and examination purposes.

It is further recommended that the multiple-choice type of test, on its own, can be used as a highly effective external moderating device where large groups of students or pupils are to be examined.

SINOPSIS

Die doel van hierdie ondersoek was om die geldigheid en betroubaarheid van twee tegnieke om skryfvaardigheid te bepaal, te evalueer. Beide metingstegnieke, nl. objektiewe (veelkeusige) en subjektiewe (opsteltipe) vrae het voor- sowel as nadele en tot dusver is daar nog nie tot 'n vergelyk met betrekking tot hul geldigheid en betroubaarheid gekom nie.

Daar word tans groot druk uitgeoefen op onderwysliggame om toe te sien dat toetse regverdig is teenoor diegene wat dit aandurf en die pleidooi vir eenvormige standaarde word steeds sterker.

Die metode van ondersoek het 'n oorsig van relevante literatuur sowel as 'n empiriese ondersoek (die toepassing van 'n Standerd 8 Engels Tweede Taal Itembanktoets en die skryf van 'n opstel en brief deur elke toetsling) ingesluit. Die toetslinge se eindeksamenpunte en hulle rapportpunte is van hulle skole verkry vir gebruik as addisionele kriteria en ook om 'n bydrae te lewer tot die statistiese analise.

'n Beskrywing van die statistiese tegnieke asook die metodes waarvolgens die geldigheid en betroubaarheid van die vier veranderlikes (die veelkeusige antwoordtipe-toets, opsteltipe-toets, die eindeksamenpunt en rapportpunt) bepaal is, word gegee.

Die resultate van die statistiese analise dui op 'n beduidende korrelasie tussen die objektiewe en subjektiewe metingstegnieke.

Die bevindinge van hierdie studie dui sterk daarop dat objektiewe en subjektiewe meting van skryfvaardigheid dieselfde grondliggende vaardighede toets. Hoewel die gesigsgeldigheid en geloofwaardigheid van objektiewe metingstegnieke deur taalonderwysers en opvoedkundiges in die algemeen bevraagteken word, word duidelik aangetoon dat die veelkeusige antwoordtipe-toets 'n geldige en betroubare instrument is om skryfvaardigheid te toets.

Die moontlikheid om objektiewe en subjektiewe evalueringsmetodes te kombineer (in een meting) en slegs die totale punt van so 'n evaluasie te vermeld, kan in aanmerking geneem word vir toets- en eksamineringsdoeleindes.

Daar word verder aanbeveel dat die veelkeusige antwoordtipe-toets, op sy eie, 'n hoogs effektiewe modereringsinstrument kan wees wanneer groot groepe studente of leerlinge geëksamineer moet word.

CONTENTS PAGE

SYNOPSIS/SINOPSIS

LIST OF TABLES

1. BACKGROUND 1

1.1 Introduction 1
1.2 Aim of this investigation 3

2. METHOD OF INVESTIGATION 4

2.1 Plan and procedure 4
2.2 Construction of measuring instruments 4

2.2.1 The multiple-choice test 4

(1) Description of content 4
(2) Examples of items 7

2.2.2 The writing test 8

(1) The choice of topics 8
(2) Topics for writing samples 9

2.2.3 Examination marks 10
2.2.4 Report marks 10

2.3 The sample 11
2.4 Method of application 14

2.4.1 The multiple-choice test 14
2.4.2 The writing test 15
(1) Testing procedure 15
(2) Scoring: (a) The scorers 15
(b) Scoring procedure 16
(c) The scoring guide 17
2.4.3 Content and method of classroom or school examinations 17

3. ANALYSIS OF THE DATA 19

3.1 The variables used in this investigation 19
3.2 Statistical properties of the multiple-choice items 19

3.2.1 Discrimination indices 20
3.2.2 Difficulty indices 21

3.3 Statistical properties of the variables 23

3.3.1 Test 1 (Multiple-choice test) 23
3.3.2 Test 2 (Writing test) 24

(1) Results of each scorer (a) Test 2 (a) (The essay) 24
(b) Test 2 (b) (The letter) 25
(2) Writing test scores 26

		PAGE
3.3.3	Examination and report marks	27
3.4	Reliability	28
3.4.1	Test 1 (Multiple-choice test)	28
3.4.2	Test 2 (Writing test)	29
	(1) Scorer reliability	29
	(2) Intercorrelations	30
3.5	Validity	
3.5.1	Content validity	32
	(1) Test 1 (Multiple-choice test)	32
	(2) Test 2 (Writing test)	33
3.5.2	Construct validity	34
3.5.3	Criterion-related validity	35
4.	IMPLICATIONS OF THE INVESTIGATION	37
4.1	Summary	37
4.2	Findings and conclusions	37
4.3	Recommendations	39
BIBLIOGRAPHY		42
APPENDIX 1:	INSTRUCTIONS FOR MARKING OF ESSAY AND LETTER	45
APPENDIX 2:	SCHOOLS INVOLVED IN THE WRITING PROFICIENCY INVESTIGATION	48

PAGE

LIST OF TABLES

TABLE 2.1 A SUMMARY OF THE STANDARD 8 ENGLISH SECOND LANGUAGE SYLLABUS, YEAR AND EXAMINATION MARKS AND THE ITEM BANK TEST SPECIFICATION TABLE 6

TABLE 2.2 THE POPULATION OF STANDARD 8 ENGLISH SECOND LANGUAGE PUPILS IN THE RSA AND THE SAMPLE INVOLVED IN THIS INVESTIGATION 13

TABLE 3.1 A DESCRIPTION OF THE VARIABLES USED IN THIS INVESTIGATION 19

TABLE 3.2 THE FREQUENCY OF DISCRIMINATION INDICES (D) OF THE ITEMS IN TEST 1 (MULTIPLE-CHOICE TEST) (TEST 1(a) COMPREHENSION AND 1(b) WRITING SKILLS AND LANGUAGE USAGE) N = 300 20

TABLE 3.3 THE FREQUENCY OF DIFFICULTY INDICES (p) OF THE ITEMS IN TEST 1 (MULTIPLE-CHOICE TEST) (TEST 1(a) COMPREHENSION AND (b) WRITING SKILLS AND LANGUAGE USAGE) N = 300 23

TABLE 3.4 THE THEORETICAL AND OBTAINED MEANS AND STANDARD DEVIATIONS (SD) OF TEST 1 (MULTIPLE-CHOICE TEST) N = 300 23

TABLE 3.5 THE MEANS, STANDARD DEVIATIONS, MINIMUM AND MAXIMUM SCORES AS ALLOCATED BY FOUR SCORERS FOR TEST 2(a) (THE ESSAY) N = 300 24

TABLE 3.6 THE MEANS, STANDARD DEVIATIONS AND MAXIMUM SCORES AS ALLOCATED BY FOUR SCORERS FOR TEST 2(b) (THE LETTER) N = 300 26

TABLE 3.7 THE MEAN AND STANDARD DEVIATION (SD) OF TEST 2 (THE WRITING TEST) N = 300 26

TABLE 3.8 THE MEANS, STANDARD DEVIATIONS, MINIMUM AND MAXIMUM SCORES OF TEST 3 (THE END-OF-YEAR EXAMINATION) AND TEST 4 (REPORT MARKS) N = 300 27

TABLE 3.9 THE RELIABILITY COEFFICIENTS AND STANDARD ERROR OF MEASUREMENT OF TEST 1 (MULTIPLE-CHOICE TEST) N = 300 28

TABLE 3.10 THE MEANS AND VARIANCES OF THE SCORES FOR TEST 2 (THE WRITING TEST) AS ALLOCATED BY THE FOUR SCORERS N = 300 29

TABLE 3.11 THE CORRELATION COEFFICIENTS BETWEEN THE SCORERS OF TEST 2(a) (THE ESSAY) N = 300 30

TABLE 3.12 THE CORRELATION COEFFICIENTS BETWEEN THE SCORERS OF TEST 2(b) (THE LETTER) N = 300 31

TABLE 3.13 THE CORRELATION COEFFICIENTS BETWEEN THE SCORERS OF TEST 2 (THE WRITING TEST) N = 300 31

TABLE 3.14 THE CORRELATION COEFFICIENTS OF SCORES OF TEST 1 (MULTIPLE-CHOICE TEST), TEST 2 (WRITING TEST), TEST 3 (EXAMINATION MARKS) AND TEST 4 (REPORT MARKS) N = 300 35

1. BACKGROUND

1.1 Introduction

In contemporary society more and more people are becoming actively concerned about the educational opportunities available to them and their children for, in order to participate constructively in their communities, a sound education has become of vital importance.

As most goals we strive to achieve require education as a means of reaching them, it is not strange that people all over the world should insist not only upon a good educational system but also upon fair and dependable measures of testing knowledge, achievement and proficiency (Stufflebeam & Shinkfield, 1985).

Tyler (1986) states that in America people stimulated by 'such things as Civil Rights movements and the promises of democratic leaders to press for educational opportunity' (p.12) have put pressures on authorities to provide quality education.

As a result, great pressure is being put on educational bodies and departments to ensure that tests are fair to all those who attempt them, that tests measure what has been taught and that results reveal to the general public what pupils can and cannot do (Tyler, 1974; Dixon, 1981; Torrance, 1986).

Because of an apparent deterioration in the writing proficiency of pupils and students, much research has recently been conducted towards the improvement as well as assessment of writing proficiency (Carlson, 1982; Sanacore, 1980; Brown & Bailey, 1984; Scherer, 1985).

Assessment is associated with written language from the early stages of acquisition, throughout the school years and for some, even through a lifetime. Spoken discourse is very seldom exposed to such close examination as is written discourse, for a writer's written attempt is much more open to scrutiny and assessment because it is in a permanent form and not as ephemeral as speech (Martlew, 1983).

When writing is, therefore, being assessed, a theoretical problem of measurement becomes one of practical importance, for one of the most important questions being asked is how a pupil can demonstrate whether a standard of proficiency in writing has been met (Prater, 1984).

Much of the research in the assessment of writing proficiency today focuses on the controversy over the use of objective type tests versus the written essay or composition. Prater points out that so far there has been little agreement as to what constitutes 'good' writing and that the issue of measuring the writing task or skill is made even more difficult by the fact that the writing skill encompasses many diverse elements.

The earliest attempts at measuring students' writing proficiency were mainly concerned with the writing skills of native speakers of English, as demonstrated by the traditional essay (Zughoul & Kambal, 1983). Although the essay test is the object of much criticism, no other form of testing seems to be as fair and as convincing to teachers, prospective employers and the general public as the actual samples of students' writing under test conditions (Diederich, 1974; Prater & Padia, 1983). Over the years critics have found it easy to demonstrate how subjective and notoriously unreliable scorers of essay tests have been. Even Hartog's classic study which pinpoints some glaring examples of the inconsistency and conflicting opinions of scorers, has been unable to dethrone the essay when used as an instrument of measurement in public examinations and other forms of testing (Huddleston, 1954).

The introduction of more objective types of testing at the turn of the century, e.g. those tests by C.W. Stone in 1908, was an attempt to replace the essay as a measuring technique.

As a result, objective tests (often standardized) mushroomed into existence in the United States of America, were marketed enthusiastically and were often used indiscriminately. A wide-spread reaction against these objective tests followed and critics pointed to

inadequate research, inadequate norms and validity studies. Their criticism resulted in a joint committee of three professional associations developing a set of formal standards for judging these objective tests and test manuals that were so readily available (Ebel, 1972, p. 458; Wolf, 1987). A set of Mental Measurement Yearbooks was introduced by Oscar Krisen Buros in 1938 to keep test users aware of the limitations as well as strengths of the tests they intend using (Ahmann, 1962).

In a country with many different cultures and group interests, such as this, setting standards in educational evaluation presents numerous problems. As pupils advance through the various school levels the demand for a common standard becomes greater. At the school leaving stage, matriculation certificates from all education departments must essentially have the same meaning, i.e. represent the same standard of achievement to future employers and to institutions of tertiary education (Malherbe, 1977).

It is especially in the second language situation, where effective techniques of measuring are a prerequisite to the improvement of language teaching and a sine qua non in an age when so many people have to study a language that is not their native tongue and where they have to demonstrate their proficiency in order to compete in an ever-growing battle to find employment.

In 1948 the De Villiers Commission on Technical and Vocational Education recommended the preparation of a series of scholastic achievement tests suitably scaled and standardized on a national basis in order that examination scores in the Republic of South Africa (RSA) could be comparable (Malherbe, 1977). The reluctance of educators and educational bodies to implement new techniques is clearly illustrated by the fact that steps in the direction of objective testing for examination purposes were only taken in 1975 with the establishment of item banks at the Human Sciences Research Council (HSRC) at the request of the Committee of Educational Heads. Thus far only the Transvaal Education Department has used multiple-choice tests from these item banks as a moderating device for their final matriculation examination.

1.2 Aim of this investigation

Over the years various researchers (Scannell & Marshall, 1966; Breland & Gaynor, 1979; Prater & Padia, 1983) have indicated several problems that occur in objective as well as subjective measuring techniques, e.g. although multiple-choice tests of writing ability may be highly reliable, they lack face validity, and measurements provided by writing samples can be unreliable because the scoring of writing samples is subject to various effects that attenuate reliability.

Hogan and Mishler (1980) state that questions concerning the relationship between objective tests and essay tests are becoming more critical in the light of demands for competency testing and criticisms of all objective tests used to assess writing skills.

Because teachers regard the essay as the best method of testing writing proficiency this investigation was primarily aimed at establishing the relationship between the 1984 Item Bank Tests for Standard 8 English Second Language and the writing samples of the Standard 8 pupils involved in this experiment.

In view of the above-mentioned controversy concerning objective versus subjective measures, the importance of investigating the usefulness of the objective test as a valid and reliable test of writing proficiency in the RSA is, therefore, self-evident.

2. METHOD OF INVESTIGATION

2.1 Plan and procedure

The plan for this investigation called for the application of an experimental multiple-choice (Item Bank) test form and the writing of an essay and a letter by a national sample of Standard 8 English second language pupils. The end-of-year examination marks as well as report marks of the pupils involved in the sample were obtained from their schools to be used as additional criterion variables.

As the main purpose of this investigation was to establish the validity of a multiple-choice test as a test of writing proficiency, the scores obtained in the multiple-choice test were to be correlated with other scores that can be regarded as more direct measures of writing proficiency.

The multiple-choice test, the writing test, the sample of testees and the procedures for administering and scoring of the tests are described in this section. The analysis of the data is reported in Section 3.

2.2 Construction of measuring instruments

2.2.1 The multiple-choice test

All the test items were of the multiple-choice type with four options each.

(1) Description of content

* Reading comprehension: Passage 1 (13 items)
  Passage 2 (15 items)

  These items included questions on content, style, inference, words in context, true/false statements, the main idea and paraphrasing.

* Vocabulary: (6 items)

  These included completion items and substituting underlined words and phrases.

* Writing skills: (7 items)

  These included questions on paragraph construction, the correct style of a letter and other editing tasks, for example spelling and punctuation.

* Language usage: (34 items)

  As editing skills are an important aspect of writing, the constructor of the test included 19 items designed to test a pupil's editing ability. These items test the correct usage of tenses, concord, prepositions, pronouns, passives and word order. (Although spelling and punctuation can be regarded as editing skills, they are dealt with under writing skills.)

The remaining 15 items were mostly completion items (pupils had to select a word or words to complete a given sentence). These include items on the correct usage of different parts of speech, prepositional phrases, word formation, direct and indirect speech, negatives, questions and synthesis.

As can be seen from Table 2.1 the multiple-choice test does not test oral or writing skills directly. However, the four language modes are not independent of each other but each contains certain linguistic aspects also found in the other. When marks are allocated for oral examinations, for example, a certain percentage will be awarded for correct usage. The same applies when writing is tested directly. Most education departments and teachers make use of a scoring guide when marking writing samples. In any scoring guide a certain percentage is allocated for correct usage, punctuation and spelling.

TABLE 2.1

A SUMMARY OF THE STANDARD 8 ENGLISH SECOND LANGUAGE SYLLABUS, YEAR AND EXAMINATION MARKS AND THE ITEM BANK TEST SPECIFICATION TABLE

The syllabus (year and examination marks)	Marks	Specification table	
1. Reading		Reading comprehension	
* A comprehension test or tests ... in contemporary English of approximately 300-400 words ... (Words in length; main points; fact/opinion; literal/figurative language; finding appropriate meanings; factual/emotive language; techniques of persuasion)	(30-35)	Passage 1: Passage 2: Content: Style: Inference: Words in context: True/false: Main idea: Paraphrasing: TOTAL	 7 2 8 6 2 1 1 28
* Literature study (No formal examination but a written examination immediately after the study of a book has been completed ... will consist of two contextual comprehension questions on each book)	60		
2. Oral		Writing skills	
* Oral work will be evaluated on the basis of continuous assessment and a year mark must be derived (prepared and unprepared speech; unprepared reading aloud with comprehension)	60	* Punctuation * Vocabulary, idiom * Correct style of a letter; addressing a person * Spelling * Paragraph construction TOTAL	2 4 2 3 3 13
3. Writing			
* To punctuate skilfully * Due attention must be given to: choice of word and idiom; sentence construction; topic sentences; paragraphing (written composition: long piece and short piece)	90		
4. Language usage		Language usage	
* Parts of speech: Nouns, prepositions, adverbs, etc.		Common errors: * Parts of speech	 8
* Tense		* Tense	9
* Concord		* Concord	3
* Voice		* Voice	2
* Direct and reported speech		* Direct and reported speech	2
* Word order		* Word order	3
* Forms of negation; emphasis		* Forms of negation	1
* Interrogative forms		* Interrogative forms	1
* Parts of a sentence		* Synthesis; misrelated clauses	3
* Word formation; spelling	(55-60)	* Word formation	2
QUARTERLY EXAMINATION MARKS (Divided proportionately as 1-4 above)	100	TOTAL GRAND TOTAL	34 75
TOTAL	400	English Second Language (Education Bulletin,1986)	
REDUCED TO GRAND TOTAL	300		

(2) Examples of items

The passages selected for comprehension purposes are regarded as tests of proficiency in understanding and interpreting English narrative and expository prose. The passages are of the type usually given to Standard 8 English second language pupils for examination purposes. As these passages and items are used for the item banks they are confidential and cannot be disclosed.

For the sections on writing skills and language usage the following item types were used (these were not taken from the test but represent items similar to those used in the test).

(a) Paragraphing: the paragraph organization type of item requires the pupil to reassemble four scrambled sentences into a coherent paragraph and then to indicate which sentence would best be the introductory or concluding sentence of this paragraph, for example:

* IF THE FOLLOWING SENTENCES WERE TO BE ARRANGED TO FORM A PARAGRAPH, WHICH SENTENCE SHOULD COME *FIRST*?

A. Another way is to soak the herbs in alcohol to make a solution.
B. For centuries herbs have been used to cure illnesses and promote good health.
C. One way is to pour boiling water over the dried leaves of the herb to make a tea.
D. There are different ways to prepare herbs for these purposes.

(b) Vocabulary items are mostly completion items or finding the best word for the underlined word, for example:

* WHICH WORD IS *OPPOSITE IN MEANING* TO THE UNDERLINED WORD?

I find such language <u>repulsive</u>.

A	B	C	D
charming	generous	liberal	elegant

(c) Language usage: these items are either in completion form or require pupils to recognise faulty usage, for example:

* WHICH WORD *BEST* COMPLETES THE SENTENCE?

During the holidays I ... travel by train.

A	B	C	D
already	usually	nearly	quickly

* WHICH UNDERLINED WORD OR WORDS ARE *WRONG*?

Men <u>have always dreamed</u> (A) of being able <u>to fly</u> (B) like birds. During the Middle Ages many would-be aeronauts <u>attached</u> (C) wings to their arms, flapped madly, and <u>have hurled</u> (D) themselves from cliffs and hills.

* WHICH UNDERLINED WORD IS SPELT *WRONGLY*?

<u>Rubbish</u> (A) is used as a <u>valueble</u> (B) raw <u>material</u> (C) in the <u>production</u> (D) of energy.

The tenses, passive forms, concord and other language structures, tested in this way, are the areas usually tested in traditional examination papers in schools.

2.2.2 The writing test

(1) The choice of topics

The traditional method of assessing writing proficiency is to present pupils with an assortment of essay, letter and paragraph topics, some easier and some more difficult than others. Pupils are allowed to make their own choice of these tasks.

It is often claimed that more difficult topics are provided for the *"benefit"* of the better pupil. This pupil, for various reasons, usually penalizes himself by choosing the more difficult task, as Britton et al. (1966) have found in a comprehensive study. Candidates who chose the more difficult tasks gained exam marks that compared less favourably with their all-round assessment marks than did the rest of the sample.

Lewis (1974) claims that when a testee is presented with a choice of questions (topics) he is, in fact, answering different papers, so the total mark may not represent comparable performances.

The constructor, therefore, limited the choice of topics (giving all testees an equal opportunity) by presenting them with one essay and one letter topic only. These topics were general and open-ended, allowing testees enough scope to use their imagination and to write to the best of their abilities.

(2) Topics for writing samples

Although the ideal situation would have been to stimulate different *"modes"* of discourse (for example descriptive or expository) to give a clearer indication of what writing abilities the testees possessed, the constructor decided to offer one narrative topic only for the essay question. (This could also eliminate the element of luck which may favour some candidates if presented with a selection of topics.)

The constructor regarded this topic as most suitable as Standard 8 English second language pupils usually have to write about their own experiences, for example an incident at school or during the holidays. The topic *"Looking over the garden wall, I saw ..."* left a wide-open spectrum of responses. The brighter pupils could opt for an imaginative, descriptive essay whilst the weaker pupils could simply describe a straightforward, unimaginative occurrence (70 marks).

The friendly letter also gave enough scope to enable all pupils to write about something familiar (a holiday job) (30 marks).

2.2.3 Examination marks

All education departments prescribe certain requirements to which an internal examination for Standard 8 pupils has to conform. An example of such a table is as follows:

* FINAL EXAMINATION

(i)	Composition	60
(ii)	Other piece of writing	20
(iii)	Comprehension	30
(iv	Language usage	30
	SUBTOTAL	140
*	PRESCRIBED READING	60
	TOTAL	200

(TED syllabus for Standard 8, 1986, p. 18)

When the above table of final examination marks is reviewed it seems that 40 per cent of the marks are allocated for the direct testing of writing. For the rest, the mode of answering even short questions requires the construction of an answer. The pupils, therefore, have to produce answers, not select them. These facts suggest that the final end-of-year examination in comparison with the multiple-choice test is more a test of writing ability. Most marks received from the schools in the sample were out of 200. For purposes of this investigation all these marks were converted to percentages.

2.2.4 Report marks

The various departmental syllabuses prescribe a system of continuous evaluation for every aspect of language teaching and the recommended

ratio of continuous assessment to the final promotion mark is 50 : 50. The following is presented as a guide for arriving at a promotion mark:

Year Mark

*	Oral	
	Listening	20
	Speaking (Prepared)	20
	(Unprepared)	20
	Reading (Prepared)	20
	(Unprepared)	20
	SUBTOTAL	100
*	Examinations/tests	100
	TOTAL	200
*	Final examination	200
	GRAND TOTAL	400

* Promotion or report marks 50 : 50

(TED syllabus for Standard 8, 1986, p. 18)

The year mark is largely based on language production skills and even for the oral section pupils are required to produce well-constructed sentences. They are, in all the above sections, required to produce rather than select answers.

These four variables, (multiple-choice test, writing test, examination marks and report marks) as described above, were used for this investigation.

2.3 The sample

For the experimental application of 7 Item Bank tests in English Second Language for Standard 8 in 1984 the Institute for Statistical Research

(HSRC) was requested to draw a random sample of 3 500 English second language pupils (500 pupils per test form being regarded as sufficient according to the formula provided by Krejcie and Morgan, 1970, p. 608). The Institute for Statistical Research selected 35 schools in the RSA in a random sample based on data supplied by the various education departments in 1981. (The universe of schools that offered English Second Language as a subject was first stratified by dividing schools into subgroups: provinces (Transvaal, Cape, Orange Free State, Natal); city/rural and boys/girls. A proportional random sample was then drawn and test material sent to 35 schools (Table 2.2). It must be noted that item bank tests for various school subjects are applied on a yearly basis. This creates organizational problems as very often a school has to apply three or more different item bank tests in one year. In order to avoid these problems the various education departments and the HSRC agreed that the testing of only two subjects per school would be allowed per year. Hence the fact that only 35 schools could be accommodated in the above sample.

Although the application of the experimental test forms was primarily undertaken for purposes of item analysis and selection for the item banks, one of these tests was to be used for the purposes of this investigation. Therefore, a special request was sent to those schools drawn in the sample to have all pupils involved in the experimental application of the item bank tests, write a letter and essay. (See Appendix 1).

These schools were also requested to submit the testees' end-of-year examination marks and final report marks to be used as external criteria.

After the experimental application and the item analysis, Test Form C was selected for purposes of this investigation as this form had the highest number of acceptable items.

It was found that 451 pupils had completed Test Form C but that not all schools had co-operated where the rest of the instructions were concerned. Some schools had sent only writing samples, others only marks and no writing samples, and so on.

For this investigation a sample of 300 testees was chosen from the original sample. This number is regarded as adequate and reasonable for obtaining the necessary statistics (Nunnally, 1967, p. 242). The number of pupils also had to be reduced to a minimum feasible level in order to save costs related to the marking of the writing samples. The composition of the sample of testees answering the multiple-choice test and writing test is set out in Table 2.2.

TABLE 2.2

THE POPULATION OF STANDARD 8 ENGLISH SECOND LANGUAGE PUPILS IN THE RSA AND THE SAMPLE INVOLVED IN THIS INVESTIGATION

Provinces		Transvaal	Cape	Natal	OFS	TOTAL
Population:		26 831	10 769	2 437	5 072	45 109
Multiple-choice test and Writing test:	*e*	178	72	16	34	300
	o	174	58	35	33	300
	x^2	0,09	2,72	22,56	0,03	25,4
x^2 at 5 % level		3,841	3,841	3,841	3,841	7,815

* *e* = the expected frequency
o = the observed frequency
x^2 = chi-square

On looking at the pupil populations of the various provinces (Table 2.2), the Transvaal has the highest number of pupils, the Cape second highest, and so on. This is also reflected in the sample of testees involved in this study as planned.

From inspection of the actual differences between observed and expected frequencies within each cell, these differences are clearly small except in the case of Natal. At the 0,05 level the difference between the observed and expected frequencies for the Transvaal, Cape and Orange Free State is not significant and it seems that the number of pupils in these cells would be regarded as proportional to the number of pupils in

their provinces. There is, however, a significant difference in the case of Natal because the province was over-represented in the sample and this difference is also reflected in the total.

Although the chi-square test for the sample over all four provinces indicates a statistically significant difference (which means that the difference can be regarded as real, rather than attributable to chance), it may be questioned whether this difference is of any practical significance as it is only the smallest province that has been over-represented by 19 pupils. Such a small number of pupils can be expected to have little impact on the total sample of 300 testees.

2.4 Method of application

2.4.1 The multiple-choice test

The Item Bank Tests in English as a Second Language for Standard 8 were applied by senior teachers or vocational guidance teachers to all the Standard 8 Higher Grade pupils in those schools that had been selected for the experimental application.

All answer sheets were returned to the HSRC for statistical processing.

2.4.2 The writing test

(1) Testing procedure

The same sample of Standard 8 second language pupils that had answered the item bank tests had to write the essay test (preferably on the day following the application of the item bank tests). The testees were allowed one and a half hours in which to complete the test.

Foolscap sheets were supplied for use and the testees were instructed to use Side 1 for the essay and Side 2 for the letter.

To avoid any more disruption (such as moving all Standard 8 pupils once more to the school hall) the testees were allowed to remain in their own classrooms. The teacher or invigilator had to write the two topics on the black board.

The essays and letters were collected afterwards, put into folders and returned together with the Item Bank answer sheets to the HSRC.

(2) Scoring

(a) The scorers

In order to facilitate the scoring of these writing samples and to obtain scores that were as objective and reliable as possible, the test constructor selected four experienced and practising teachers from the Transvaal Education Department (two of whom regularly act as Assistant Chief Markers for the Standard 10 final examination). One teacher from a rural area was able to attend the discussions on the marking procedure and formed part of the panel of four. For practical reasons teachers from other provinces could not be invited to join the panel.

(b) Scoring procedure

These four scorers attended a meeting in which the scoring guide (see Appendix 1) was explained. The scorers were asked to follow this guide carefully, first reading rapidly for a total impression and then reading more slowly for an evaluation of the language aspect.

It was emphasized that the scorers should not be excessively influenced by language errors, but rather rely on their global impression of the writing sample - whether a testee could communicate and convey his ideas even if the language used was grammatically poor.

After the first ten essays and letters had been scored, the test constructor controlled the marks awarded by each scorer and discussed any difficulties or problems regarding the scoring guide. The different marks awarded by the scorers were discussed and an average mark given to each of the essays and letters scored during this session. These essays and letters were to be used as models for future reference.

The above procedure was repeated twice (five essays and letters) to assist the scorers in maintaining standards.

One scorer was inclined to penalize too heavily for poor grammar and was advised to lay more emphasis on content and whether the testee was able to communicate.

Each scorer received 300 copies of the original writing samples and each writing sample was scored individually by the four scorers.

There was no indication from which province or school the writing samples had been selected as each school was known only by a number allocated to it earlier. The scorers could, therefore, not be influenced by a foreknowledge of which particular school or area they were assessing.

(c) The scoring guide

Nitko (1983) suggests that holistic scoring is probably more appropriate when essays involving a testee's abilities to create and to synthesize are being scored (especially where no definite correct answer is prespecified). The constructor, therefore, requested the scorers to take the various aspects associated with the scoring of writing samples (for example grammar, punctuation and word order), into consideration, but to rely to a great extent on their holistic or global impression when awarding marks. The allocation of definite sets of marks for the various aspects of writing samples may result in a distortion in the assessment of its value, especially if too much emphasis is laid on correct grammatical usage (see Appendix 1).

2.4.3 Content and method of classroom or school examinations

In spite of their importance as a measuring instrument neither the end-of-year examination mark nor the report mark can be regarded as very reliable criteria. There are various factors that contribute to the unreliability and invalidity of these marks.

Firstly, tests and examination papers are compiled by different teachers for different classes in different schools. These teachers could all be aiming at measuring the same objectives in their tests but, because of their individuality, could be placing more weight on one aspect of content to be tested than another, thus producing different papers with possibly different standards of difficulty. This means that these tests or examination papers are nonparallel.

Secondly, the suitability of examination questions is not determined empirically (questions are seldom, if ever, pre-tested for difficulty or discrimination). No "*internal consistency or other form of reliability is published*" (Lewis, 1974, p. 91). Furthermore, teachers are allowed a free hand in their choice of questions.

Thirdly, when these papers are scored they are scored by individual teachers who do not always maintain the same standard of scoring. The decisions of a teacher very often reflect his knowledge and reasoning thereby influencing the teacher to set subjective standards based on what pupils (in the opinion of the teacher) should attain.

Fourthly, when term or year marks have to be determined, the teacher, himself, usually decides which scores (oral, writing samples and language tests) to combine for a final mark. Each of these components also carries more or less weight, according to the judgment of the individual teacher.

In defence of teacher-made tests and their scores, Ingram (1977) claims that experienced teachers are capable of ranking the members of their classes in order of merit in respect of their relative achievement in a school subject, and that scores produced by teachers appear to be one of

the most widely available criterion measures. It is also important to remember that there "*is no means of test validation that is completely empirical, is completely impersonal and objective, and avoids the vagueness and uncertainty of human judgments altogether*" (Ebel, 1972, p. 442).

Nitko (1983) comments that almost all criterion measures are incomplete and "*represent only part of the attainment of the ultimate behaviour series*" a test ought to predict (p. 424).

Although test and examination scores can be regarded as contaminated with irrelevant factors, they remain attractive as criteria, for they are usually easily obtainable, readily quantifiable and of great importance in making decisions such as pass/fail and the placement of pupils. The marks obtained from the schools in the sample were to be used as additional criteria for this investigation.

3. ANALYSIS OF THE DATA

3.1 The variables used in this investigation

In order to elucidate the content of each variable, the variables are described in Table 3.1

TABLE 3.1

A DESCRIPTION OF THE VARIABLES USED IN THIS INVESTIGATION

TEST 1	The multiple-choice test	1(a) Comprehension items (b) Writing skills and language usage
TEST 2	The writing test	2(a) Essay (b) Letter
TEST 3	The end-of-year examination marks	Marks for composition, another piece of writing, comprehension, language usage and literature
TEST 4	Final report marks	A combination of marks given during the year and final examination marks

3.2 Statistical properties of the multiple-choice items

Of the total number of items in Test Form C, 76 items had discrimination indices higher than 0,200 (75 per cent of the items). Of these 76 items, 75 items were selected for Test 1 (the multiple-choice test) as one item, although it had a satisfactory discrimination index, was very difficult (below 20 per cent).

These 75 items were then divided into two subgroups (those items based on comprehension passages and the others concerned with writing skills and language usage) and item analysis was carried out once more. Only the answer sheets of those 300 pupils involved in the writing proficiency study were scored. This above division was made in order to obtain more information concerning those skills being tapped by comprehension passages and skills involved in writing and language usage and how these subtests correlated with the other variables used in this investigation.

After the second computer processing programme it was found that these 75 items fell well within the criteria set out below and that there was no need to discard any more items. Table 3.2 indicates the distribution of item discrimination indices of Test 1 (the comprehension, writing skills and language usage items) after the deletion of items that did not discriminate adequately.

3.2.1 Discrimination indices

The index of discrimination is a most important criterion when the acceptability of an item as a whole has to be established (Nunnally, 1970). The point-biserial correlation with the total score of the test is used in this investigation and the criterion level of correlations above 0,20 has been adopted for the index of discrimination for Test 1 (Nunnally regards correlations above 0,20 as acceptable).

TABLE 3.2

THE FREQUENCY OF DISCRIMINATION INDICES (D) OF THE ITEMS IN TEST 1 (MULTIPLE-CHOICE TEST) (TEST 1(a) COMPREHENSION AND 1(b) WRITING SKILLS AND LANGUAGE USAGE) N = 300

Test	0,200 - 0,299	0,300 - 0,399	0,400 - 0,499	0,500 - 0,599	Total number of items	Mean (D)
Test 1(a)	10	8	9	1	28	0,351
Test 1(b)	18	18	8	3	47	0,337
Test 1(a) + 1(b)	28	26	17	4	75	0,344

The mean of the discrimination indices of the revised test (Test 1) is 0,344 which, according to Ebel (1972) presents a fairly satisfactory picture.

3.2.2 Difficulty indices

An essential requirement of the correct response is that the group of testees that selected this response should obtain a higher average in the test than those testees that selected any of the alternative responses (Chamberlain, 1973, p. 124). Should this not be the case the item must be revised or rejected.

The number of testees selecting the correct response (expressed as a percentage of those who attempted the item) indicates the difficulty of that item for that particular group (the higher the percentage the easier the item).

Ebel (1972, p. 390) states that the difficulty values of items need not be considered "*directly*" in item selection as no item which is much too difficult or much too easy can possibly discriminate well.

Nunnally (1967, p. 273) adds that when item analysis of achievement tests is executed more stress should be laid on correlations of items with the total scores than their difficulty.

However, Chamberlain and Gericke (1983) observe that it has been determined both theoretically and empirically that the most reliable and valid scores can be obtained by means of items of medium difficulty. In the case of multiple-choice items the term "*medium*" does not equal 50 per cent but has to be adjusted for chance level, as indicated below:

Number of options per item	Theoretical mean as percentage
2	75
3	66
4	62½
5	60

Fairly wide deviations from the theoretical mean can still be regarded as "*medium*" difficulty, as indicated below:

Number of options per item	Range of "medium difficulty" as percentage
2	60-90
3	47-87
4	40-85
5	36-84

According to Chamberlain (1973) Hardaway suggests that the index of difficulty can range between 10 per cent and 90 per cent. In the case of four options this means that chance level would allow for a difficulty level of 100 - 7½% = 92½ per cent for the top cut-off point and for the bottom cut-off point as 25 + 7½. This allows for 32½ per cent as the lowest index of difficulty.

The test constructor decided to adopt a more lenient criterion and to lower the index of difficulty to 25 per cent as a few items in the experimental test form still discriminate very well in spite of a low difficulty value.

The range of the index of difficulty of those items selected from Form C would, therefore, be between 25 per cent and 92½ per cent.

Although provision had been made for a top difficulty value of 92,5 per cent, it was found that no item had a difficulty index of more than 85 per cent or less than 25 per cent after the elimination of items that did not discriminate adequately. (This was after the second computer processing programme in which the 75 items selected for Test 1 had been involved.)

Table 3.3 indicates the frequency of the difficulty indices of the items in the revised test (Test 1).

TABLE 3.3

THE FREQUENCY OF DIFFICULTY INDICES (p) OF THE ITEMS IN TEST 1 (MULTIPLE-CHOICE TEST) (TEST 1(a) COMPREHENSION AND (b) WRITING SKILLS AND LANGUAGE USAGE) N = 300

Test	20.1-30.0	30.1-40.0	40.1-50.0	50.1-60.0	60.1-70.0	70.1-80.0	80.1-90.0	Total	Mean
Test 1(a)	2	1	4	3	6	10	2	28	61,5
Test 1(b)	2	3	11	10	8	10	3	47	57,5
Test 1(a) + 1(b)	4	4	15	13	14	20	5	75	59,6

The mean of the difficulty indices of the 75 items in Test 1 is 59,6 per cent. (The significance of this figure will be discussed in 3.3.1).

3.3 Statistical properties of the variables

3.3.1 Test 1 (Multiple-choice test)

TABLE 3.4

THE THEORETICAL AND OBTAINED MEANS AND STANDARD DEVIATIONS (SD) OF TEST 1 (MULTIPLE-CHOICE TEST) N = 300

Test	Theoretical mean		Obtained mean		Difference		Theoretical SD	Obtained SD	Difference
	Raw score	Percentage	Raw score	Percentage	Raw score	Percentage			
Test1	46,8	62,5	44,4	59,6	2,4	3,2	9,37	12,93	3,56

The theoretical mean of a test can be regarded as a score value corresponding to 50 per cent of the total score of the test (Owen, 1971). Should this criterion be adopted, it is preferable that the obtained mean of the test should not differ by more than 5 per cent from the theoretical mean (Chamberlain, 1973) and that the maximum deviation of the obtained mean from the theoretical mean should not exceed 10 per cent.

The theoretical mean of a test consisting of 4-choice items is estimated at 62,5 per cent (Chamberlain & Gericke, 1983). The mean of Test 1 is satisfactory as it differs by only 2,47 from the theoretical mean of 46,87 for 75 items. There is, however, an indication that the test could possibly have been slightly difficult for the testees as the mean of Test 1 is 2,47 lower than the theoretical mean (raw scores).

Both the mean and the standard deviation of Test 1 are satisfactory. (The mean is very close to the midpoint between chance level and the maximum possible score).

The skewness (SK) of Test 1 is 0,04 which is a low positive value and indicates that the distribution of this test is very close to that of a normal distribution. The kurtosis is 2,31 which indicates a flatter score distribution than for the normal kurtosis which has a value of 3.

3.3.2 Test 2 (Writing test)

(1) Results of each scorer

(a) Test 2(a) (The essay)

The statistical properties of Tests 2(a) and 2(b) as scored by four scorers were calculated separately. The means, standard deviations, minimum and maximum scores for Test 2(a) are indicated in Table 3.5.

TABLE 3.5

THE MEANS, STANDARD DEVIATIONS, MINIMUM AND MAXIMUM SCORES AS ALLOCATED BY FOUR SCORERS FOR TEST 2(a) (THE ESSAY) N = 300

Scorers	Mean	SD	Minimum	Maximum	Total
Scorer 1	30,1	10,31	5	58	70
Scorer 2	35,2	10,83	8	64	70
Scorer 3	37,5	11,32	13	65	70
Scorer 4	36,2	9,32	16	64	70

If the theoretical mean of a test is regarded as a score value corresponding to 50 per cent of the total score of the test, the theoretical mean for Test 2(a) should be 35. From the above it becomes clear that Scorer 2 was closest to the theoretical mean and Scorer 4 second closest.

Scorer 1 marked more strictly and Scorer 3 was most lenient although closer to the theoretical mean than Scorer 1.

As already stated in 2.4.2 (2)(b) one of the scorers (during the training period) was inclined to score more analytically (possibly still affected by the rigidity of departmental scoring guides which, in many instances, expect scorers to add up the number of language errors and then to place the writing sample in a particular error category, thus restricting the number of marks the

testee could have obtained if the writing sample had been marked more holistically). The minimum and maximum marks awarded by Scorer 1 are also lower than those of the other scorers.

On looking at Scorers 2, 3 and 4 the means of their essay scores are very close: 35,2; 37,5 and 36,2 respectively, indicating that they had tuned in well to the method laid out in the general scoring guide (Appendix 1). The same applies for the maximum scores awarded. The theoretical SD of the essay test, Test 2(a), should be 10. Scorer 4 was the only scorer whose scores had a lower SD than that of the theoretical SD. The standard deviations of the other three scorers were above the theoretical SD which makes them more acceptable. (Correlations between the scorers will be discussed in 3.4.2).

(b) Test 2(b) (The letter)

In Table 3.6 the statistical properties of Test 2(b) as scored by the four scorers are indicated.

TABLE 3.6

THE MEANS, STANDARD DEVIATIONS AND MAXIMUM SCORES AS ALLOCATED BY FOUR SCORERS FOR TEST 2(b) (THE LETTER) N = 300

Scorers	Mean	SD	Minimum	Maximum	Total
Scorer 1	13,1	4,37	1	27	30
Scorer 2	16,5	4,31	6	26	30
Scorer 3	16,0	4,26	7	27	30
Scorer 4	17,3	3,90	8	26	30

In the case of the letter sample (Test 2(b)) the theoretical mean should be 15 and, in this case, Scorer 3 is closest to the theoretical mean with 16,0.

The scores of Scorer 1 once more indicate the lowest mean. Scorers 2, 3 and 4 were in close agreement on the minimum scores and all four scorers were close when awarding the maximum scores.

The theoretical SD should be 4,3 and only Scorers 1 and 2 are above the theoretical SD. However, all four scorers are satisfactorily close to the theoretical SD. (The correlations for the combined test score will be discussed in 3.4.2).

(2) Writing test scores

In Table 3.7 the mean and standard deviation of Test 2 (the essay and letter samples) are indicated.

TABLE 3.7

THE MEAN AND STANDARD DEVIATION (SD) OF TEST 2 (THE WRITING TEST) N = 300

Test	Theoretical Mean (Percentage)	Obtained Mean (Percentage)	Difference	Theoretical SD	Obtained SD	Difference
Test 2	50	50,4	0,4	10,40	11,91	1,46

It seems that the mean of Test 2 (the essay and letter writing samples) is very close to the theoretical mean of 50 per cent, only 0,4 above. The SD of Test 2 is also satisfactory as it is 1,46 above the theoretical SD of 10,4.

In conclusion, the mean score obtained by the testees was very close to the theoretical mean which indicates that Test 2 was scored appropriately for the ability level of the group of testees.

The skewness (SK) for Test 2 is 0,13 which indicates that the distribution of test scores is somewhat positively skewed. A kurtosis of 2,73 is indicated which means that the distribution of scores is slightly flatter than that of a normal curve distribution.

3.3.3 Examination and report marks

In Table 3.8 the statistics for the end-of-year examination and final report marks are set out.

TABLE 3.8

THE MEANS, STANDARD DEVIATIONS, MINIMUM AND MAXIMUM SCORES OF TEST 3 (THE END-OF-YEAR EXAMINATION) AND TEST 4 (REPORT MARKS) N = 300

Test	Mean	SD	Minimum	Maximum	Total
Test 3 (Examination marks)	56,2	13,69	17	89	100
Test 4 (Report marks)	56,9	12,43	25	88	100

The means for Tests 3 and 4 are close and are much higher than the theoretical mean of 50 per cent (but still satisfactory). The standard deviations for these tests are also satisfactory. The SK for the score distribution of Test 3 (end-of-year examination marks) is 0,10. The SK for the score distribution of Test 4 (report marks) is 0,20. Both these test score distributions are positively skewed. The score distribution of Test 3 has a kurtosis of 2,73, which is the same as that of Test 2. The score distribution of Test 4 has a kurtosis of 2,61, which indicates a slightly flatter distribution than those of Tests 2 and 3.

3.4 Reliability

3.4.1 Test 1 (Multiple-choice test)

The K-R 8 and K-R 21 formulas were applied to establish the reliability coefficients of Test 1, the multiple-choice test.

Table 3.9 indicates the K-R 8 and K-R 21 reliability coefficients as well as the standard error of measurement of the raw scores of the subtests.

TABLE 3.9

THE RELIABILITY COEFFICIENTS AND STANDARD ERROR OF MEASUREMENT (SEM) OF TEST 1 (MULTIPLE-CHOICE TEST) N = 300

Test	Number of items	K-R 8	K-R 21	SEM (Raw scores)
Test 1(a)	28	0,85	-	2,12
Test 1(b)	47	0,88	-	2,91
Test 1(a) + 1(b)	75	-	0,90	4,09

The K-R 8 formula for Test 1(a) indicates a reliability coefficient of 0,85. For Test 1(b) the K-R 8 indicates a reliability coefficient of 0,88. This is to be expected as Test 1(b) has a larger number of items.

The K-R 21 formula was used to estimate the reliability of the multiple-choice test as a whole, as the reliability coefficient for the test total was not provided by the item analysis programme. (This formula gives a computationally simpler approximation to the K-R 20.)

The reliability coefficient for Test 1 as a whole is 0,90. Popham (1981) regards 0,80 as adequately reliable, especially where groups are being tested, and in relation to this criterion, the reliability indices of these subtests and test as a whole can be regarded as satisfactorily high.

3.4.2 Test 2 (Writing test)

There are two types of indices which are helpful in judging the reliability of the writing test:

* Scorer reliability (the extent to which the scorers agree with one another) and

* the intercorrelations of the scores of the two writing tasks.

(1) Scorer reliability

To enhance the reliability of the essay scores, four experienced scorers (who scored all the writing tasks independently) were used for this study. The correlation between the scores of the different scorers for a particular writing task provides a measure of the reliability with which these writing tasks were read.

Coefficient alpha (α), was used to estimate the reliability of the essay test. (The correlation coefficient serves as an index of the extent to which the scorers were consistent.) Coefficient alpha is interpreted in essentially the same way as the K-R 20 formula (Nitko, 1983).

Table 3.10 indicates the means of the marks assigned by the various scorers.

TABLE 3.10

THE MEANS AND VARIANCES OF THE SCORES FOR TEST 2 (THE WRITING TEST) AS ALLOCATED BY THE FOUR SCORERS N = 300

Variable	N	Maximum	Mean	SD
Scorer 1 (Essay, Letter)	300	100	43,17	13,7
Scorer 2 (Essay, Letter)	300	100	51,75	14,3
Scorer 3 (Essay, Letter)	300	100	53,43	14,2
Scorer 4 (Essay, Letter)	300	100	53,55	12,0
Essay, Letter	300	400	201,24	

After having applied the coefficient alpha formula, it was found that the reliability coefficient of the essay (Test 2(a)) was 0,86, for the letter (Test 2 (b)) 0,89, and for Test 2 as a whole, 0,90. This indicates that the reliability of these scores was satisfactorily high, in fact, appreciably higher than the inter-scorer correlations discussed in the following section. This also supports the conclusion that the multiple-marking of the writing

test makes it a better criterion of writing proficiency than the end-of-year examination or report mark where only one scorer is used and where other aspects of achievement in English are tested in addition to writing proficiency.

(2) Intercorrelations

The Pearson Product-moment correlation coefficient was used to determine the inter-scorer reliability of the two writing tasks (the essay and the letter) separately. Tables 3.11, 3.12 and 3.13 respectively indicate the correlation coefficients of the essay (Test 2(a)), the letter (Test 2(b)) and Test 2 (a + b) as scored by the four scorers.

TABLE 3.11

THE CORRELATION COEFFICIENTS BETWEEN THE SCORERS OF TEST 2(a) (THE ESSAY) N = 300

	Scorer 1	Scorer 2	Scorer 3
Scorer 1	1,000		
Scorer 2	0,705		
Scorer 3	0,546	0,538	
Scorer 4	0,685	0,692	0,535

From the above table it becomes clear that the scores of Scorer 1 correlate reasonably high with those of Scorers 2 and 4 respectively (0,705 and 0,685). The scores of Scorer 3, however, have the lowest correlation coefficients with the scores of Scorers 1, 2 and 4 respectively (0,546, 0,538 and 0,535) although these correlations are still in the reasonable category (according to Mulder, 1981). The scores of Scorers 2 and 4 also correlate satisfactorily (0,692). According to Guilford (1965) reliability coefficients are expected to be in the upper brackets of r values, usually 0,70 to 0,98.

TABLE 3.12

THE CORRELATION COEFFICIENTS BETWEEN THE SCORERS OF TEST 2(b) (THE LETTER) N = 300

	Scorer 1	Scorer 2	Scorer 3
Scorer 1	1,000		
Scorer 2	0,730		
Scorer 3	0,562	0,563	
Scorer 4	0,721	0,693	0,552

It seems that the letter scores of the four markers have slightly higher correlations than those of the essay scores although the scores of Scorer 3 seem (in both cases) to have the lowest correlation indices (0,562, 0,563 and 0,552).

TABLE 3.13

THE CORRELATION COEFFICIENTS BETWEEN THE SCORERS OF TEST 2 (THE WRITING TEST) N = 300

	Scorer 1	Scorer 2	Scorer 3
Scorer 1	1,000		
Scorer 2	0,773		
Scorer 3	0,602	0,595	
Scorer 4	0,749	0,752	0,609

Table 3.13 indicates the correlation coefficients of Test 2 (writing test) as scored by the four scorers. The same pattern as indicated in Table 3.11 and 3.12 is repeated. However, the correlation coefficients for the test as a whole are higher. The scores of Scorer 1 correlate reasonably highly with those of Scorers 2 and 4 respectively (0,773 and 0,749). The scores of Scorer 3 have the lowest correlation coefficients with the scores of Scorers 1, 2 and 4 respectively (0,602, 0,595 and 0,609) although they are higher than in the subtests and fall in the high category (according to Mulder, 1981). This is in agreement with the Spearman-Brown formula that tests containing more items are more reliable.

3.5 VALIDITY

3.5.1 Content validity

(1) Test 1 (Multiple-choice test)

To determine the content validity of the multiple-choice test (Test 1) the content of the test was specified and compared to the type of language test usually set for Standard 8 English Second Language pupils. (See Table 2.1) Since all possible items in a language test domain cannot be included in a single test, this multiple-choice test can only depend on the representativeness of its content.

Test 1 was based on the assumption that the ability of a pupil to handle language in the reading mode would closely correspond to his command of writing and would, therefore, constitute a valid means of testing a pupil's writing proficiency.

In order to set a valid multiple-choice test which is also a representative sample of the universe of language skills advocated in the syllabus, the following procedure was followed:

(a) The universe of content was divided into a number of subcategories:

* Reading comprehension.

* Writing skills: paragraphing, mechanics, spelling and punctuation.

* Language usage: use and abuse of words, direct and indirect speech, passives, word order.

(b) A relative representation of each category was decided upon (the importance of weighting was taken into consideration); and

(c) items were then selected from each category until the number of items decided upon in 3.5.1(2) had been reached (Chamberlain, 1973).

As Test 1 was approved by a committee of subject experts and the items written by experienced teachers, it can be assumed that Test 1 is a valid sample of the whole domain (See Table 2.1). (Content validity cannot be determined statistically but is based on the expertise of qualified people (Mulder, 1981).)

(2) Test 2 (Writing test)

The writing test has always been regarded as a valid criterion as it represents a direct measure of writing and also propounds the type of writing task, for example narrative or descriptive, as stated in the core syllabus for Standard 8 English Second Language pupils. However, the number of writing tasks as well as the sampling of tasks is greatly restricted in a test or examination situation owing to a lack of time. The latter fact inhibits the content validity of the writing test.

In a language and writing test, pupils are usually presented with a list of 5 or 6 topics (these are usually expected to be in the narrative, expository or descriptive mode). The pupils are then to write on one topic only. In the same way they have to choose a letter topic from a list of two or three topics.

Sometimes pupils may choose between a letter or a dialogue. (This constitutes the general form of a Standard 8 English Second Language writing test as advocated in the core syllabus.) In the case of Test 2, all the pupils had to write an essay on one topic only and a letter on one topic only. Although both Test 1 and Test 2 can be accepted as reasonable tests in terms of content validity, Test 2 seems to have more face validity.

3.5.2 Construct validity

Construct validity need not necessarily be verified by statistical means only but may also be determined logically (Mulder, 1981). The hypothe-

sis that the various criteria, for example Test 1 (the multiple-choice test), should be related to the construct of writing proficiency could be accepted when a high correlation between Test 2 (the writing test) which is regarded as the most direct way of measuring writing proficiency, (Lewis, 1974; Breland & Gaynor, 1979) and Test 1 (the multiple-choice test) is indicated.

In this case, the correlation of 0,752 between Test 2 and Test 1 indicates that Test 1 (the multiple-choice test) possesses a fairly high degree of construct validity and could be regarded as a reasonably valid test of writing proficiency. (Construct validation by means of a factor analysis programme was not attempted for this investigation).

The hypothesis that there would be a higher correlation between Test 1 and Test 2 than between Test 1 and Tests 3 and 4 respectively was tested by means of Hotelling's formula for a t-test of the significance of the difference between correlations of dependent variables. In this formula the number of degrees of freedom may be taken as N - 3 (Guilford, 1965, p. 190).

The t-value for the difference between the correlation coefficient of Test 1 (the multiple-choice test) and Test 2 (the writing test) on the one hand and the correlation coefficient of Test 1 and Test 3 (end-of-year examination marks) on the other, is 1,09.

The t-value for the difference between the correlation coefficient of Test 1 and Test 2 on the one hand and the correlation coefficient of Test 1 and Test 4 (report marks) on the other, is 0,81.

The above indicates that there is only a slight difference between the correlations involved.

When the 0,05 level of significance is adopted, the t-value needed is 1,64 for a one-tailed test of significance. It can, therefore, be concluded that the true difference in correlations is not significant at the 0,05 level of probability.

3.5.3 Criterion-related validity

The results discussed in this section highlight the empirical findings concerning the two variables: the multiple-choice test and the writing test.

The criterion-related validity of Tests 1 and 2 can be determined when the correlation coefficients between the testees' writing samples (the essay and letter), the multiple-choice test, the end-of-year examination and the report marks respectively are computed.

The correlation coefficients of Test 1 (the multiple-choice test) and Test 2 (the writing test) are indicated in Table 3.14.

Table 3.14 indicates the correlation coefficients for all the variables used in this study.

TABLE 3.14

THE CORRELATION COEFFICIENTS OF SCORES OF TEST 1 (MULTIPLE-CHOICE TEST), TEST 2 (WRITING TEST), TEST 3 (EXAMINATION MARKS) AND TEST 4 (REPORT MARKS) N = 300

TEST 1(a) (Comprehension skills)	TEST 1(a)	TEST 1(b)	TEST 1	TEST 2	TEST 2(a)	TEST 2(b)	TEST 3
TEST 1(b) (Writing skills & Language usage)	0,720						
TEST 2 (Writing Test)	0,652	0,731	0,752				
TEST 2(a) (Essay)	0,630	0,706	0,726				
TEST 2(b) (Letter)	0,593	0,687	0,698		0,756		
TEST 3 (Examination)	0,635	0,697	0,722	0,700	0,661	0,657	
TEST 4 (Report)	0,678	0,722	0,757	0,733	0,693	0,674	0,904

The correlation coefficient for Test 2(a) and Test 2(b) is 0,756 which is reasonably high. This confirms the construct validity of the two parts of the writing test. The correlation coefficient for Test 1 and Test 2(a) is 0,726 (still reasonably high) and 0,698 for Test 1 and Test 2(b).

The correlation coefficient for Test 1 and Test 2 is 0,752 which is reasonably high. It can, therefore, be deduced that a high concurrent validity exists between the multiple-choice test (Test 1) and the writing test (Test 2).

In fact, this correlation is slightly, but not significantly, (at the 0,01 level of probability) higher than the correlation between two reliable measures of writing proficiency (the correlations between Tests 2(a) and 2(b) mentioned in the previous paragraph).

As the examination marks (Test 3) are incorporated in the final report marks (Test 4), there is a spuriously high correlation between Tests 3 and 4, as indicated by 0,904.

The correlation coefficient for Test 2 and Test 3 is reasonably high: 0,700. So, too, is the correlation between Test 2 and Test 4 which is 0,733. There is no statistically significant difference between these two correlations at the 0,01 level of probability.

The correlation coefficient for Test 1 and Test 3 is 0,722 and for Test 1 and Test 4, 0,757.

It is important to note that the intercorrelation indices are all in the high category, indicating that Test 1 (the multiple-choice test) possesses a high criterion-related validity and can be regarded as a useful measure for assessing writing proficiency (correlation with Test 2) and achievement in English (correlations with end-of-year examination and report marks).

The experimental investigation reported in the preceding sections has, therefore, established the fact that the multiple-choice test is satisfactorily effective for predicting scores on a reliable criterion of writing proficiency. Furthermore, it is at least as effective a predictor of end-of-year results in English as is a highly reliable writing test.

4. IMPLICATIONS OF THE INVESTIGATION

4.1 SUMMARY

The importance of obtaining evidence concerning the reliability and validity of multiple-choice tests, especially when being used for examination purposes, can best be comprehended against the background of the profound changes that have taken place in education measurement. A new awareness of the complexity of education, the problems concerning fair examinations and the promotion of new technologies have strongly influenced developments in educational measurement. The arrival of the computer has also introduced a vast new range of possibilities for the recording and storing of data in a way that allows almost instantaneous retrieval and processing.

As great pressure is being put on educational bodies to ensure quality education and fair examinations (as discussed in 1.1), this investigation has been an attempt to make some contribution to this problem area by comparing two methods of assessing writing proficiency.

As described in Section 2, the method of this investigation involved an empirical study entailing the application of experimental item bank tests and the writing of an essay and a letter by each testee in the nation-wide sample.

4.2 FINDINGS AND CONCLUSIONS

An analysis of the empirical data rendered the following results:

The sample

In spite of every effort to obtain the writing test scores, examination and report marks of all the pupils involved in the experimental application of the Item Bank test, this was not possible. As discussed in 2.3, it was found that the actual differences between observed and expected frequencies within each cell, with the exception of Natal, were small. At the 0,05 level the differences between the observed and the expected frequencies for the Transvaal, Cape and Orange Free State were not significant. These samples could, therefore, be regarded as representative of their provinces. It was only in the case of Natal that there was a significant difference between the observed and the expected frequencies because of that province being over-represented in the sample. It may be questioned whether the latter fact could be of any practical significance as only a small number of pupils were involved in this province in relation to the total sample.

The means and standard deviations

As discussed in 3.3, the means and standard deviations of the variables involved in this investigation were all satisfactory.

Correlation coefficients

On looking at the correlation coefficients, Test 1 (the multiple-choice test) correlated highly with all the other variables. The correlation between Test 1 and Test 4 (Report marks) was only slightly higher than that between Test 1 and Test 2 (Writing test).

* From the findings above, it can be concluded that there is no empirical reason why a multiple-choice test cannot be used as a test to measure writing proficiency. This does not mean that the writing process should be neglected. Writing should continue to be taught and tested in the classroom.

* The writing test is a relatively efficient method of testing small groups. When testing on a nation-wide basis, however, various

weaknesses of the writing test as a measuring instrument are highlighted. As discussed in 3.3.2, it is difficult to key in scorers to score writing samples, especially when they have to remain consistent and score reliably. It is important, too, to have multiple-scoring of writing samples in order to obtain the highest reliability. In a nation-wide examination (even on a provincial level) the cost involved will be prohibitive.

On comparing the correlations between the two variables (the multiple-choice test and the writing test), it can be concluded that the multiple-choice test is at least as good a measure of writing proficiency as is the highly valid and reliable measure of writing proficiency - the writing test used in this investigation.

4.3 RECOMMENDATIONS

Good and effective methods are essential to the control of teaching and learning and there is a need for continual assessment of the progress of education. Any measuring device must be sound and the outcome of tests or examinations must be valid and reliable.

* For curriculum development

Although this investigation has indicated the suitability and efficacy of a multiple-choice test as a means of assessing writing proficiency, this by no means negates the importance of testing writing by means of writing samples. Writing will always remain an important skill in a literate society and should, therefore, remain an ongoing process in the classroom.

* For test construction

Although the Item Bank Tests for Standard 8 English Second Language pupils were originally intended to test the language proficiency of these pupils and not specifically their writing proficiency, the results clearly indicate that some of the skills involved in the writing process were also being tapped. A multiple-choice test, therefore, which was specifically constructed to test writing proficiency and which consisted of more items based specifically on

such writing skills as mechanics, paragraph construction and other editing skills, could possibly yield even higher reliability and validity indices.

* For test development

As revealed by the results (discussed in 3.3, 3.4 and 3.5) the multiple-choice test has indicated its *suitability as a reliable and valid means of testing writing proficiency*. This is significant too, in that it confirms what researchers such as Godshalk et al. (1966); Breland and Gaynor (1979) and Karzmarek (1980) have already reported in their various studies.

The results of this investigation are especially meaningful as they imply that a multiple-choice test can act as a highly effective external moderating device where large groups of pupils or students are to be tested. In addition, the relatively low cost involved in the selecting of items for a test from the HSRC Item Bank and the scoring of such a test, can be highly advantageous.

It should also be kept in mind that the validity and reliability of current matriculation papers are questionable where heterogeneous groups and test items that have not previously been tested, are involved. It is almost an impossibility to compare the various examination papers throughout the country and to establish whether they have the same standards. It is therefore recommended that item bank tests be used to equate the results of different examining bodies.

* For the development of item banks

Not all the item banks at the HSRC are, as yet, fully functional or computerized and in some subjects only a limited number of items are available owing to syllabus changes. The ideal situation would be for schools to be able to draw multiple-choice tests from the

bank for their own use. This, however, is too costly. At this stage, the item banks at the HSRC can supply moderating tests to examining bodies that are interested in using these tests for application to large groups of testees. The possibility of developing tests for particular groups or schools can also be considered.

* Finally, it must be emphasized that before concluding that objective multiple-choice English Second Language items focus on superficial aspects of writing only, educational authorities ought to give serious consideration to the data reported here.

Objective testing, especially by means of multiple-choice tests, still has to attain its rightful place in the RSA. Many teachers and educators need to become aware of the effectiveness and usefulness of objective measures of testing and that a well-constructed multiple-choice test can be a reliable and valid means of testing language proficiency in general and writing proficiency in particular.

Today a vast and unexplored field for research in writing lies beckoning. There is also a need for teachers and educators to come to grips with modern technology and testing methods and to use this knowledge productively for the betterment of language teaching and testing in the RSA.

BIBLIOGRAPHY

Ahmann, J.S. (1962). *Testing student achievement and aptitudes.* Washington, D.C.: The Center for Applied Research in Education.

Breland, H.M. & Gaynor, J.L. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement,* 16(2), 119-128.

Britton, J.N., Martin, N.C. & Rosen, H. (1966). *Multiple marking of English compositions: An account of an experiment.* London: H.M.S.O., 1966.

Brown, J.D. & Bailey, K.M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning,* 34 (4), 21-42.

Carlson, (1982). *Survey of ETS writing activities* (Division of Educational Research and Evaluation). Princeton, N.J.: Educational Testing Service.

Chamberlain, J.C. (1973). *The development of a proficiency test battery in Arithmetic, English and Afrikaans for Post-Form III Bantu pupils.* M.Ed. thesis. Pretoria: UNISA.

Chamberlain, J.C. & Gericke, F.W. (1983). *A manual for the planning of objective scholastic achievement tests and the writing of multiple-choice items.* Pretoria: HSRC.

Diederich, P.B. (1974). *Measuring growth in English.* Urbana, Ill.: National Council of Teachers of English.

Dixon, R.R. (1981). After the test: There should be no mystery. *New Directions for Testing and Measurement,* 9, 65-70.

Ebel, R.L. (1972). *Essentials of educational measurement.* Englewood Cliffs, N.J.: Prentice Hall.

English Second Language, Higher Grade Standards 8, 9 and 10 (1986). *Education Bulletin* (Department of Education and Culture), 9 (Special Edition), 1-21.

Godshalk, F.I., Swineford, F. & Coffman, W.E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Guilford, J.P. (1965). *Fundamental statistics in psychology and education*. New York: McGraw-Hill.

Hogan, T.P. & Mishler, C.J. (1980). Relationships between essay tests and objective tests of language skills for elementary school students. *Journal of Educational Measurement*, 17(3),219-227.

Huddleston, E.M. (1954). Measurement of writing ability at the College-entrance level: Objective versus subjective testing techniques. *Journal of Experimental Education*,22(3), 165-213.

Karzmarek, C.M. (1980). Scoring and rating essay tasks. In J.W. Oller Jr. & K. Perkins (Eds.), *Research in language testing* (pp. 151-159). Rowley, Mass.: Newbury House.

Krejcie, R.V. & Morgan, D.W. (1970). Determining sample size for research activities. *Educational and psychological measurement*, 30, 607-610.

Lewis, D.G. (1974). *Assessment in education*. London: University of London Press.

Malherbe, E.G. (1977). *Education in South Africa: Vol. II, 1923-75*. Cape Town: Juta.

Martlew, M. (1983). Cognitive and communicative aspects of writing. In M. Martlew (Ed.), *The psychology of written language* (pp. 295-333). Bath: Pitman.

Mulder, J.C. (1981). *Statistiese tegnieke in opvoedkunde*. Pretoria: HAUM.

Nitko, A.J. (1983). *Educational tests and measurement: An introduction*. New York: Harcourt Brace Jovanovich.

Nunnally, J.C. (1967). *Psychometric theory*. New York: McGraw - Hill.

Nunnally, J.C. (1970). *Introduction to psychological measurement*. New York: McGraw - Hill.

Owen, K. (1971). *Opstelling en standaardisering van die akademies-tegniese aanlegtoetse vir kleurlinge* (ATA). Pretoria: RGN.

Popham, W.J. (1981). *Modern educational measurement*. Englewood Cliffs: Prentice-Hall.

Prater, D. (1984). The development of a criterion-referenced test of basic writing skills. *Journal of Research and Development in Education*, 18 (1), 12-16.

Prater, D. & Padia, W. (1983). Developing parallel holistic and analytic scoring guides for assessing elementary writing samples. *Journal of Research and Development in Education*, 17(1), 20-24.

Sanacore, J. (1980). *School-wide strategies*. Clearing House, 53, 391-393.

Scannell, D.P. & Marshall, J.C. (1966). The effect of selected composition errors on grades assigned to essay examinations. *American Educational Research Journal*, 3, 125-130.

Scherer, D.L. (1985). *Measuring the measurements: A Study of evaluation of writing: An annotated bibliography*. Bloomington, Ind.: Indiana University. (ERIC Document Reproduction Service No. ED 260-455).

Stufflebeam, D.L. & Shinkfield, A.J. (1985). *Systematic evaluation*. BOSTON: Kluver Academic.

Torrance, (1986). What can examinations contribute to school evaluation? *Educational Review*, 38(1) 31-43.

Transvaal Education Department (1986). *Perspectives: Ordinary Grade*. Pretoria: TED.

Tyler, R.W. (1974). The use of tests as measuring the effectiveness of educational programs, methods, and instrumental materials. In R.W. Tyler & R.M. Wolf (Eds.), *Crucial issues in testing* (pp. 143-155). Berkeley, Calif.: Mc Cutchan.

Tyler, R.W. (1986). Societal changes influencing educational evaluation. *International Journal of Educational Research*, 10 (1), 11-19.

Wolf, R.M. (1987). Educational evaluation: The state of the field. *International Journal of Educational Research*, 11(1), 1-143.

Zughoul, M.R. & Kambal, M.E. (1983). Objective evaluation of EFL composition. *IRAL*, 21(2), 87-101.

APPENDIX 1: INSTRUCTIONS FOR MARKING OF ESSAY AND LETTER

1. GENERAL COMMENT ON THE EVALUATION OF WRITTEN COMPOSITION

The most reliable assessment of an essay or letter is usually based on a global impression. The allocation of definite sets of marks for the various aspects of the essay or letter may result in a distortion in the assessment of its value. For the sake of uniformity of marking, however, an analysis of the various aspects of creative writing (essay and letter) is indispensable. Although the markers have to take these aspects into consideration, they are to use their own discretion and rely to a great extent on their <u>global impression</u>.

NB Markers are therefore reminded <u>not</u> to judge these Std 8 English Second Language essays and letters chiefly on the aspects of language and usage. Aspects of content and structure should also receive consideration.

2. INSTRUCTIONS FOR MARKING OF ESSAY AND LETTER

(1) Do remember that you are evaluating the work of a candidate in a language which is not his mother tongue.

(2) Markers are required to mark all the work done by pupils. The general principle is to indicate <u>all</u> errors where practicable.

(3) Errors are indicated as follows:

(a) <u>Language usage (grammar)</u>: tense, concord, prepositions, pronouns, etc.) a circle: he (are)

(b) <u>Word use and idiom:</u> a wavy line

(c) <u>Word order</u>: an arrow

(d) <u>Omissions</u>: a caret (^)

(e) <u>Punctuation:</u> Full stop x
comma: insert ,

(f) <u>Spelling</u>: Vertical or oblique stroke: m/ny

(4) Whenever you are seriously in doubt as to what a mark should be, please indicate.

(5) No marks are deducted if an answer is longer than the stipulated length as pupils generally penalize themselves. All errors must be indicated. Do not merely 'scan' part of a too long essay, as this often causes a wrong estimate.

(6) An essay falling short of the required length should be penalized pro rata.

3. EVALUATION SCHEME FOR ESSAY AND LETTER

The procedure for awarding a content mark is as follows:

(1) Evaluate the essay as well as the letter in terms of each of the criteria below (CRIS) with reference to the Positive to Negative explanations.

(2) Then choose a category in the CONTENT column of the table below in terms of the global impression arrived at.

Positive	1 - 2 - 3 - 4 - 5	Negative
Coherence	Lucid/Logical development/ ... well organised/ Unity of structure	Garbled/confused woolly/disconnected
Relevance	Recognition of implications ... and boundaries of essay and letter topics	Avoidance or deviation from topic/ padding
Ideas	Perceptive/genuine/fresh ..... original/action/feeling Atmosphere	Contrived/dull/poverty of thought
Substantiation	Thoughts well illustrated	Insufficient

e.g. C = 3
R = 2
I = 3
S = 3
11

11
4
Content column = 3

(3) Now choose the appropriate category in the EXPRESSION COLUMN e.g. C (Essay is fairly competent but contains errors of tense and concord). Marks for this particular essay will then be anything between 35 and 41.

ESSAY	EXPRESSION				
CONTENT	A Excellent Thoroughly at home in English/Consistent excellence/Errors of no consequence	B Pleasing Signs of literary ability/ Competent/ Errors of minor consequence	C Fairly fluent Pedestrian/ Fairly competent/ Some errors in concord and tense	D Clumsy Forced/ Careless/ Slangy/ Errors disturbing	E Incoherent Faulty in idiom/Awkward sentence structure/Incompetent/Repeated errors of serious nature
CRIS 1 (Excellent)	A63-70	A56-62	B49-55		
CRIS 2 (Above average)	A56-62	B49-55	C42-48	D35-41	
CRIS 3 (Average)	B49-55	C42-48	D35-41	E28-34 (40 %)	F21-27
CRIS 4 (Below average)		D35-41	E28-34 (40 %)	FF21-F27	G20
CRIS 5 (Very weak)			FF21-27	G14-20	H0-13

LETTER	EXPRESSION				
CONTENT	A Excellent	B Pleasing	C Fairly fluent	D Clumsy	E Incoherent
CRIS 1 (Excellent)	A27-30	A24-26	B21-23		
CRIS 2 (Above average)	A24-26	B21-23	C18-20	D15-17	
CRIS 3 (Average)	B21-23	C18-20	D15-17	E12-14 (40 %)	FF9-F11
CRIS 4 (Below average)		D15-17	E12-14 (40 %)	FF9-F11	G8
CRIS 5 (Very weak)			FF9-11	G6-8	H0-5

N.B.: Lists of names are provided. Markers are requested to fill in marks given to the essays and letters in the appropriate columns.

APPENDIX 2:

SCHOOLS INVOLVED IN THE WRITING PROFICIENCY STUDY

		Number of pupils
A.	TRANSVAAL	
1.	Parkstraat Hoër Handelskool (Pretoria)	12
2.	Hoërskool Die Adelaar (Roodepoort)	15
3.	Handelskool Johan Jurgens (Springs)	12
4.	Hoërskool Louis Trichardt	10
5.	Hoërskool Edenvale (Johannesburg)	12
6.	Hoërskool Hoogland (Brakpan)	26
7.	Hoërskool Wesvalia (Klerksdorp)	13
8.	Hoërskool Westonaria	16
9.	Hoërskool Balfour	12
10.	Hoërskool Ogies	8
11.	Hoërskool John Vorster (Nigel)	14
12.	Wagpos Sekondêre skool (Brits)	10
13.	Hoër Tegniese skool (Vereeniging)	7
14.	Hoërskool Barberton	7
	TOTAL (68 boys, 106 girls)	174
B.	ORANGE FREE STATE	
15.	C en O Sekondêre Meisieskool Oranje (Bloemfontein)	15
16.	Hoër Tegniese skool Sasolburg	14
17.	Bloemspruit Sekondêre skool (Shannon)	4
	TOTAL (15 boys, 18 girls)	33
C.	NATAL	
18.	Hoërskool Ferrum (Newcastle)	10
19.	Gelofte Hoërskool (Pinetown)	16
20.	Hoërskool Suid-Natal (Port Shepstone)	9
	TOTAL (14 boys, 21 girls)	35
D.	CAPE PROVINCE	
21.	Boland Landbou Hoërskool	6
22.	Napier Hoërskool	3
23.	Vryburg Hoërskool	11
24.	Middelburg Hoërskool	5
25.	Vaalrivier Hoërskool (Barkly West)	3
26.	Hoërskool Graaff-Reinet	8
27.	Hoërskool Stella	5
28.	Port Rex Hoër Tegniese skool (East London)	6
29.	Brackenfell Hoërskool (Cape Town)	11
	TOTAL (33 boys, 25 girls)	58
	GRAND TOTAL	300

280D

HUMAN SCIENCES RESEARCH COUNCIL
RAAD VIR GEESTESWETENSKAPLIKE NAVORSING

President	Dr J.G. Garbers	President
Deputy Presidents	Dr H.C. Marais, Dr J.D. Venter	Adjunk-presidente
Vice-Presidents	Dr K.F. Mauer, Prof. D.J. Stoker	Vise-presidente
Executive Director: Administration	J.G.G. Gräbe	Uitvoerende Direkteur: Administrasie
Chief PRO	Dr G. Puth	Skakelhoof

Functions of the HSRC

The HSRC undertakes, promotes, supports and co-ordinates research in the field of the human sciences. It also determines research priorities, disseminates the findings of human sciences research, facilitates and evaluates the implementation of research findings, stimulates the training of researchers, places the full spectrum of human sciences disciplines at the service of the inhabitants of the RSA and promotes science in general.

Funksies van die RGN

Die RGN onderneem, bevorder, ondersteun en koördineer navorsing op die gebied van die geesteswetenskappe, bepaal navorsingsprioriteite, versprei die resultate van geesteswetenskaplike navorsing, vergemaklik en evalueer die implementering van die resultate van navorsing, stimuleer die opleiding van navorsers, stel die volle spektrum van dissiplines in die geesteswetenskappe ten diens van die inwoners van die RSA en bevorder die wetenskap in die breë.

Institutes

Institute for Communication Research (ICOMM)

Institute for Educational Research (IER)

Institute for Historical Research (IHR)

Institute for Manpower Research (IMAN)

National Institute for Personnel Research (NIPR)

Institute for Psychological and Edumetric Research (IPER)

Institute for Research Development (IRD)

Institute for Research into Language and The Arts (IRLA)

Institute for Sociological and Demographic Research (ISODEM)

Institute for Statistical Research (ISR)

Bureau for Research Support Services (BRSS)

Administration

Institute

Instituut vir Geskiedenisnavorsing (IGN)

Instituut vir Kommunikasienavorsing (IKOMM)

Instituut vir Mannekragnavorsing (IMAN)

Instituut vir Navorsingsontwikkeling (INO)

Instituut vir Opvoedkundige Navorsing (ION)

Nasionale Instituut vir Personeelnavorsing (NIPN)

Instituut vir Psigologiese en Edumetriese Navorsing (IPEN)

Instituut vir Sosiologiese en Demografiese Navorsing (ISODEM)

Instituut vir Statistiese Navorsing (ISN)

Instituut vir Taal- en Kunstenavorsing (INTAK)

Buro vir Ondersteunende Navorsingsdienste (BOND)

Administrasie

Head office

Private Bag X41, Pretoria 0001
Republic of South Africa
Telegrams RAGEN
Tel. (012) 202-9111
Telex 3-20893 SA

Hoofkantoor

Privaatsak X41, Pretoria 0001
Republiek van Suid-Afrika
Telegramme RAGEN
Tel. (012) 202-9111
Teleks 3-20893 SA

NIPR

P.O. Box 32410, Braamfontein 2017
Republic of South Africa
Telegrams NAVORSPERS
Tel. (011) 339-4451
Telex 4-25459 SA

NIPN

Posbus 32410, Braamfontein 2017
Republiek van Suid-Afrika
Telegramme NAVORSPERS
Tel. (011) 339-4451
Teleks 4-25459 SA

Regional offices

Western Cape, Private Bag X5, Roggebaai 8012
Tel. (021) 419-2572/3/4/5 Telex 5-22260 SA

Natal, P.O. Box 17302, Congella 4013
Tel. (031) 815970 Telex 6-28567 SA

NIPR Eastern Cape, P.O. Box 1124, Port Elizabeth 6000
Tel. (041) 53-2131 Telex 2-43203 SA

Streekkantore

Wes-Kaap, Privaatsak X5, Roggebaai 8012
Tel. (021) 419-2572/3/4/5 Teleks 5-22260 SA

Natal, Posbus 17302, Congella 4013
Tel. (031) 815970 Teleks 6-28567 SA

NIPN Oos-Kaap, Posbus 1124, Port Elizabeth 6000
Tel. (041) 53-2131 Teleks 2-43203 SA